IDICSO
Instituto de Investigación en Ciencias Sociales
Facultad de Ciencias Sociales
Universidad del Salvador

ÁREA EMPLEO Y POBLACIÓN

La regresión logística

por **Horacio Chitarroni***

Buenos Aires, DIC/2002

* **CHITARRONI, Horacio.** Lic. en Sociología, Universidad Nacional de Buenos Aires (UBA). Docente, Facultad de Ciencias Sociales, Universidad del Salvador (USAL). Docente de la Maestría en Ciencias Sociales del Trabajo, Facultad de Ciencias Sociales, UBA. Investigador Principal, Área Empleo y Población, IDICSO, USAL. Consultor del Consejo Nacional de Coordinación de Políticas Sociales, SIEMPRO (Sistema de Evaluación, Seguimiento y Monitoreo de Programas Sociales).

BREVE HISTORIA DEL IDICSO. Los orígenes del IDICSO se remontan a 1970, cuando se crea el "Proyecto de Estudio sobre la Ciencia Latinoamericana (ECLA)" que, por una Resolución Rectoral (21/MAY/1973), adquiere rango de Instituto en 1973. Desde ese entonces y hasta 1981, se desarrolla una ininterrumpida labor de investigación, capacitación y asistencia técnica en la que se destacan: estudios acerca de la relación entre el sistema científico-tecnológico y el sector productivo, estudios acerca de la productividad de las organizaciones científicas y evaluación de proyectos, estudios sobre política y planificación científico tecnológica y estudios sobre innovación y cambio tecnológico en empresas. Las actividades de investigación en esta etapa se reflejan en la nómina de publicaciones de la "Serie ECLA" (SECLA). Este instituto pasa a depender orgánica y funcionalmente de la Facultad de Ciencias Sociales a partir del 19 de Noviembre de 1981, cambiando su denominación por la de Instituto de Investigación en Ciencias Sociales (IDICSO) el 28 de Junio de 1982.

Los fundamentos de la creación del IDICSO se encuentran en la necesidad de:

- Desarrollar la investigación pura y aplicada en Ciencias Sociales.
- Contribuir a través de la investigación científica al conocimiento y solución de los problemas de la sociedad contemporánea.
- Favorecer la labor interdisciplinaria en el campo de las Ciencias Sociales.
- Vincular efectivamente la actividad docente con la de investigación en el ámbito de la facultad, promoviendo la formación como investigadores, tanto de docentes como de alumnos.
- Realizar actividades de investigación aplicada y de asistencia técnica que permitan establecer lazos con la comunidad.

A partir de 1983 y hasta 1987 se desarrollan actividades de investigación y extensión en relación con la temática de la integración latinoamericana como consecuencia de la incorporación al IDICSO del Instituto de Hispanoamérica perteneciente a la Universidad del Salvador. Asimismo, en este período el IDICSO desarrolló una intensa labor en la docencia de post-grado, particularmente en los Doctorados en Ciencia Política y en Relaciones Internacionales que se dictan en la Facultad de Ciencias Sociales. Desde 1989 y hasta el año 2001, se suman investigaciones en otras áreas de la Sociología y la Ciencia Política que se reflejan en las series "Papeles" (SPI) e "Investigaciones" (SII) del IDICSO. Asimismo, se llevan a cabo actividades de asesoramiento y consultoría con organismos públicos y privados. Sumándose a partir del año 2003 la "Serie Documentos de Trabajo" (SDTI).

La investigación constituye un componente indispensable de la actividad universitaria. En la presente etapa, el IDICSO se propone no sólo continuar con las líneas de investigación existentes sino también incorporar otras con el propósito de dar cuenta de la diversidad disciplinaria, teórica y metodológica de la Facultad de Ciencias Sociales. En este sentido, las áreas de investigación del IDICSO constituyen ámbitos de articulación de la docencia y la investigación así como de realización de tesis de grado y post-grado. En su carácter de Instituto de Investigación de la Facultad de Ciencias Sociales de la Universidad del Salvador, el IDICSO atiende asimismo demandas institucionales de organismos públicos, privados y del tercer sector en proyectos de investigación y asistencia técnica.

IDICSO

Departamento de Comunicación

Email: idicso@yahoo.com.ar

Web Site: http://www.salvador.edu.ar/csoc/idicso

Introducción

La regresión logística es un instrumento estadístico de análisis multivariado, de uso tanto explicativo como predictivo. Resulta útil su empleo cuando se tiene una variable dependiente dicotómica (un atributo cuya ausencia o presencia hemos puntuado con los valores cero y uno, respectivamente) y un conjunto de variables predictoras o independientes, que pueden ser cuantitativas (que se denominan covariables o covariadas) o categóricas. En este último caso, se requiere que sean transformadas en variables "dummy", es decir variables simuladas[1].

El propósito del análisis consiste en:

predecir la probabilidad de que a alguien le ocurra cierto "evento": por ejemplo, estar desempleado =1 o no estarlo = 0, ser pobre = 1 o no pobre = 0, recibirse de sociólogo =1 o no recibirse = 0).

Determinar que variables pesan más para aumentar o disminuir la probabilidad de que a alguien le suceda el evento en cuestión

Esta asignación de probabilidad de ocurrencia del evento a un cierto sujeto, así como la determinación del peso que cada una de las variables dependientes en esta probabilidad, se basan en las características que presentan los sujetos a los que, efectivamente, les ocurren o no estos sucesos.

Por ejemplo, la regresión logística tomará en cuenta los valores que asumen en una serie de variables (edad, sexo, nivel educativo, posición en el hogar, origen migratorio, etc.) los sujetos que están efectivamente desocupados (=1) y los que no lo están (=0). En base a ello, predecirá a cada uno de los sujetos – independientemente de su estado real y actual – una determinada probabilidad de ser desocupado (es decir, de tener valor 1 en la variable dependiente). Digamos, si alguien es un jóven no jefe de hogar, con baja educación y de sexo masculino y origen migrante (aunque esté ocupado) el modelo le predecirá una alta probabilidad de estar desocupado (puesto que la tasa de desempleo de el grupo así definido es alta), generando una variable con esas probabilidades estimadas. Y procederá a clasificarlo como desocupado en una nueva variable, que será el resultado de la predicción.

Y además, analizará cuál es el peso de cada uno de estas variables independientes en el aumento o la disminución de esa probabilidad. Por ejemplo, cuando aumenta la educación disminuirá en algo la probabilidad de ser desocupados. En cambio, cuando el sexo pase de 0 = mujer a 1 = varón, aumentará en algo la probabilidad de desempleo porque la tasa de desempleo de los jóvenes de sexo masculino es mayor que la de las jóvenes mujeres. El modelo, obviamente, estima los coeficientes de tales cambios.

[1] Se incluye un breve apéndice explicativo acerca de la transformación de variables categóricas en variables dummy.

Cuanto más coincidan los estados pronosticados con los estados reales de los sujetos, mejor ajustará el modelo.

Un ejemplo de regresión logística

Se trata de predecir la probabilidad de ser pobres de los jefes de hogar, a partir de datos de la EPH.

Variable dependiente:

- pobreza (previamente recategorizada en 0=no pobre y 1=pobre)

Variables independientes cuantitativas (covariadas):

- Cantidad de ocupados formales en el hogar
- Cantidad de asalariados no precarios en el hogar
- Cantidad de ocupados con calificación profesional en el hogar
- Cantidad de ocupados con calificación técnica en el hogar
- Clima educativo del hogar
- Promedio de edad de los miembros del hogar
- Cantidad de menores de 14 años en el hogar
- Cantidad de desocupados en el hogar
- Cantidad de jubilados en el hogar
- Cantidad de perceptores de ingresos en el hogar
- Tamaño del hogar

Variables independientes categóricas (se convierten en dummy):

- nivel educativo del jefe del hogar (se define que la base sea nivel = 1, sin instrucción)

La función logística refleja la probabilidad del evento (ser pobre), expresada como "*Odds*" (chances):

Prob pobre/prob no pobre = E^z

Donde:

E = base del logaritmo natural (2,718)

Z = a + b1X1 + b2X2 + ...bnXn

De manera que:

Z = Log. PP/PNP (Prob pobre/prob no pobre)

a = constante del modelo (como la ordenada al origen de la regresión, es decir el valor de la variable dependiente cuando todas las variables independientes son = cero)

X = variables independientes

b = pesos de cada variable independiente, que pueden ser positivos o negativos (cuando X varía en una unidad, el logaritmo del cociente PP/PNP aumenta o disminuye en b unidades)

Tablas de resultados

Uno de los primeros indicadores de importancia para apreciar el ajuste del modelo logístico es el doble logaritmo del estadístico de Likelihood. Se trata de un estadístico que sigue una distribución similar a Chi Cuadrado y compara los valores de la predicción con los valores observados en dos momentos: a) en el modelo sin variables independientes, sólo con la constante y b) una vez introducidas las variables predictoras. Por lo tanto, el valor del Likelihood debiera disminuir sensiblemente entre ambas instancias e – idealmente – tender a cero cuando el modelo predice bien. En este caso, esa primera estimación es: 4794,71

```
Beginning Block Number  0.  Initial Log Likelihood Function

-2 Log Likelihood   4794,71

* Constant is included in the model.
```

Posteriormente, se introducen las variables independientes y aparece nuevamente el valor de Likelihood: se ha reducido a 3408,22 (aproximadamente en 10%, que no es demasiado). Y se nos avisa que el modelo ha iterado (realizado rutinas de cálculo) cinco veces, hasta que la reducción resultó menor a 0,01%.

```
Estimation terminated at iteration number 5 because
Log Likelihood decreased by less than ,01 percent.

 -2 Log Likelihood     3408,218
```

Luego, tenemos otro modo de apreciar la bondad de las predicciones: una tabla de contingencia nos muestra los valores reales de pobre/no pobre que asumen los sujetos y los que les predice el modelo. Cuanto mayor sea la carga en la diagonal positiva de la tabla, mejor será el ajuste:

```
Classification Table for ESLP
The Cut Value is ,50
                    Predicted
                   ,00    1,00     Percent Correct
                     0  I    1
Observed         +-------+-------+
   ,00      0    I  1402 I   444 I   75,95%
                 +-------+-------+
   1,00     1    I   383 I  1240 I   76,40%
                 +-------+-------+
                           Overall  76,16%
```

En este caso, tenemos un 76% de predicciones correctas. Sobre 1623 personas pobres, el modelo les predijo esa condición a 1240. Con los no pobres, la situación es similar. Se nos dice también que el "punto de corte" para asignar una persona a las categorías pobre/no pobre fue de 50%: se lo clasifica como pobre cuando las chances de serlo superan ese valor. Esto se puede modificar a voluntad (por ejemplo, situarlo en 75%).

Se calcula también un chi cuadrado para esta tabla, que permite rechazar la hipótesis nula de distribución al azar con alta significación:

```
Chi-Square    df Significance
 Model                1386,492    19        ,0000
 Block                1386,492    19        ,0000
 Step                 1386,492    19        ,0000
```

La tabla logística:

Finalmente, tenemos el elemento decisivo para la interpretación del modelo: la tabla logística que proporciona los pesos y significación de cada variable en la predicción del evento, donde:

B: son las estimaciones de las b de la ecuación

S.E.: es el error estándar de estas estimaciones

Wald: el estadístico de Wald es una prueba de significación estadística que testea la hipótesis nula de que las b son iguales a cero

DF: son los grados de libertad de cada variable

Sig.: es el nivel de significación del Wald (vale decir la probabilidad de error al descartar la hipótesis nula)

R: es similar al coeficiente de correlación parcial y varía entre 1 y –1, indicando la capacidad de determinación de cada variable manteniendo el resto constantes

Exp (B): este estadístico nos dice cuanto aumenta (o disminuye) el "Odds ratio", o sea el cociente Pp/pnp (luego de que X aumenta en una unidad) / sobre

Pp/pnp (antes de que X varíe). El valor 1 indica que la variable no influye. Valores superiores a 1 indican aumento y valores inferiores a 1 indican disminución.

Lower y Upper (que aparecen en algunas versiones recientes del SPSS) son los límites del intervalo de confinza donde se sitúa el verdadero valor de Exp (B), con 95% de probabilidad. Por lo tanto, si este intervalo de confianza pasa por 1, no podemos descartar la hipótesis nula de que estos "odds ratio" no varíen al variar X.

---------------------- Variables in the Equation ----------------------

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
FORMAL_1	-,5963	,0956	38,8833	1	,0000	-,0877	,5508
PRECAR_1	-,2512	,0976	6,6287	1	,0100	-,0311	,7778
CALIPR_1	-,8818	,3562	6,1297	1	,0133	-,0293	,4140
CALITE_1	-,3462	,1541	5,0501	1	,0246	-,0252	,7074
CLIMAEDU	-,3244	,0237	187,8293	1	,0000	-,1969	,7229
EDAPROME	-,0319	,0053	36,1933	1	,0000	-,0844	,9686
ZMEN14	,1864	,0586	10,1304	1	,0015	,0412	1,2049
DESOCU_1	,4762	,0887	28,8052	1	,0000	,0748	1,6099
JUBIPE_1	-,0474	,1292	,1343	1	,7140	,0000	,9537
PERCEP_1	-,1354	,0507	7,1443	1	,0075	-,0328	,8734
POBTOT	,0512	,0374	1,8684	1	,1717	,0000	1,0525
NIVEL			8,9222	8	,3489	,0000	
NIVEL(1)	,2703	,1667	2,6283	1	,1050	,0114	1,3104
NIVEL(2)	,2070	,1649	1,5761	1	,2093	,0000	1,2300
NIVEL(3)	,2976	,1643	3,2822	1	,0700	,0164	1,3467
NIVEL(4)	,0894	,1978	,2043	1	,6513	,0000	1,0935
NIVEL(5)	,7456	,4057	3,3781	1	,0661	,0170	2,1077
NIVEL(6)	-,1055	,4451	,0562	1	,8126	,0000	,8998
NIVEL(7)	-,0387	,2718	,0203	1	,8867	,0000	,9620
NIVEL(8)	-,0958	,5224	,0337	1	,8544	,0000	,9086
Constant	3,2634	,3714	77,2265	1	,0000		

Dependent Variable.. ESLP

Un primer criterio para apreciar los resultados es ceñirse a aquellas variables para las cuales el estadístico de Wald arroja significación. Ello permitiría desechar el tamaño del hogar, la presencia de jubilados y casi todos los niveles educativos del jefe, a excepción de los niveles 3 y 5 (significativos al 10%: probabilidad de error de tipo I cercana a 7%).

Luego, para las variables que resultan significativas (es decir, donde la probabilidad de que b sea cero en la población es muy reducida), podemos examinar los valores y el signo de estos coeficientes b:

por cada ocupado formal "adicional" en el hogar se reduce en 0,6 el logaritmo de la chance de ser pobre por cada asalariado precario el logaritmo de esta chance disminuye levemente, en 0,25 en cambio, la presencia de un trabajador con calificación profesional disminuye muy fuertemente la chance de ser pobre: su logaritmo se reduce en 0,88 el aumento del clima educativo del hogar

reduce el logaritmo de las probabilidades de ser pobre: casi en un tercio por cada año por cada desocupado en el hogar, en cambio, el logaritmo de la chance de pobreza se incrementa en 0,47 (aquí b es positivo) los niveles crecientes de educación del jefe reducen, como sería de esperar, las chances de pobreza (signos de b negativos), pero no arrojan significación (no podemos descartar que esos b sean cero en la población).

Los coeficientes R, tal como la correlación parcial, nos dicen cuanto pesa cada variable en el aumento o disminución de la probabilidad del evento, con prescindencia de las otras. Así, podemos ver que la presencia de trabajadores formales en el hogar reduce la probabilidad de ser pobre. Igualmente, pero con un peso considerablemente mayor lo hace el clima educativo del hogar. En cambio, esta probabilidad aumenta si hay desocupados en el hogar.

Finalmente, los Exp (b) nos indican la relación entre las chances pobre/no pobre, "antes y después" de que cada variable independiente aumente en una unidad. Cuando ese coeficiente es 1, entonces quiere decir que no cambian esas chances. Si Exp (b)es mayor a uno, entonces esas chances aumentan. Por el contrario, cuando es menor a uno, ellas disminuyen. Por ejemplo, el aumento de un ocupado formal hace que se reduzca este cociente a casi la mitad (supongamos que el cociente prob pobre/prob no pobre fuera 0,70 antes de aumentar un ocupado formal: luego del aumento habría pasado a ser 0,35).

Como producto de la aplicación del modelo se crean variables nuevas:

- La clasificación pobre/no pobre según la predicción
- La probabilidad asignada por el modelo en términos de "odds" (pp/pnp)

Apéndice

Las variables "dummy"

A veces necesitamos incorporar al modelo de regresión logística variables independientes que no son numéricas sino categóricas. Supongamos, por ejemplo, que queremos predecir la probabilidad de ser pobre de una persona. Tal vez nos resulte importante incorporar variables que no son cuantitativas: por ejemplo, la categoría ocupacional (empleador, cuentapropista, asalariado, trabajador sin remuneración). En este caso, esta variable podría ser incorporada a la ecuación si se la transforma en una variable *dummy* (simulada). Ello consiste en generar n – 1 variables dicotómicas con valores cero y uno, siendo n el número de categorías de la variable original.

Para el caso de la variable categoría ocupacional, la transformación sería la siguiente:

Categoría ocupacional	Variables dummy		
	Empleador	Cuenta propia	Asalariado
Empleador	1	0	0
Cuenta propia	0	1	0
Asalariado	0	0	1
Trabajador sin remuneración	0	0	0

Crearíamos tres variables dicotómicas: la primera de ellas sería "Empleador". Quien lo sea tendrá valor 1 en esa variable y valor cero en las variables "Cuenta propia" y "Asalariado". Los cuentapropistas tendrán valor 1 en la segunda variable y cero en las otras, etc. No necesitamos crear, en cambio, una variable llamada "Trabajador sin remuneración": lo será quien tenga valores cero en las tres anteriores. Esta última es la categoría "base" de las *dummy*[2].

Una vez realizada esta transformación, estas variables pueden ser incorporadas en una ecuación de regresión: sus valores sólo pueden variar entre cero y uno[3] y sus coeficientes b indicarán, en cada caso, cuanto aumentan o disminuyen los "odds" de probabilidad del evento que se procura predecir cuando una de estas variables pasa de cero a uno (por ejemplo, cuando alguien es un

[2] Obviamente, podríamos haber definido como base cualquiera de las cuatro categorías.
[3] Al haber un solo intervalo, no puede haber intervalos desiguales. Son, pues, "variables de intervalos iguales".

empleador, seguramente la probabilidad de que sea pobre disminuirá, lo que se expresará en un coeficiente b negativo en la ecuación logística).

Buenos Aires, DIC/2002

por **Horacio Chitarroni**
Investigador Principal, Área Empleo y Población, IDICSO, USAL.
Email: hchitarroni@siempro.gov.ar